INTRODUCTION TO MEASURES OF CENTRAL TENDENCY AND VARIABILITY - Part I.

George Geologist was given the assignment of determining the lithologic characteristics of the Ying-Yang Formation. Five outcrops of the complete unit were measured along FM 123.4 and thicknesses of green mudstone, quartz arentite, biomicrite and red mudstone are recorded in Table 1.1.

Table 1.1

Outcrop	Green Mudstone	Quartz Arenite	Biomicrite	Red Mudstone
1	20'	30'	40'	50'
2	10'	15'	15'	30'
3	5'	10'	30'	20'
4	15'	20'	20'	25'
5	20'	20'	25'	10'

A working familiarity with the concepts of matrix algebra is becoming increasingly important in the geosciences and will be discussed in class. Table 1.1 is a matrix (T); each object (in this case an outcrop) constitutes a row of the matrix and the measurements on that object are given in the columns of the matrix. An individual piece of information (an element of the matrix) is specified by giving its row and column location. For example, T(3,2) is the entry in the third row and second column - 10' of quartz arenite at outcrop 3. In general, T(i,j) is the entry in the ith row and jth column.

George hands in his table and impatiently awaits his grade. His advisor returns the table with a big **F** on it and asks the following questions:

(1) what are you trying to determine?
(2) why did you select these 5 outcrops?
(3) what is the average amount of the 4 lithologies?
(4) which lithology is the most variable?
(5) do you know what you are doing?

Questions like these come up all the time during a study. Unfortunately, for George, these questions came after considerable effort had been spent in making measurements. Some of these questions are best considered before data are collected, others during the study and still others only after the study is completed. Central to many investigations is the distinction between the **target** and **sampled populations**. The target population is the set of all conceivable observations of a particular type. The investigator selects the target population prior to data collection. For example, George's advisor was interested in lithologic thickness variations in the Ying-Yang Formation. He could narrow the target population by specifying only subsurface samples or only outcrops or only published accounts, etc. Restrictions placed on the limits of a target population may make the results of the investigation of limited value; for example, all outcrops within a five minute walk of a road might be easy on George but poor on science. For many target populations it may be physically impossible, too expensive, or too time consuming to examine all of the potential members of the population and the investigator must construct a sampled population which is a subset of the target population. The sampled population must be defined in such a way that all members of the target population have an equal chance of being included in the sampled population. Information based on observations from the sampled population may be used (with proper care) to make statements about the target population.

Statistics can be thought of as describing various properties of the sampled population. **Parameters** describe properties of the target population. Following convention, small Greek letters will be used to denote parameters and small English letters for statistics.

Univariate Statistics - Describing the Columns in a Data Table

Having made a set of measurements you will probably want to say something about their average value. The arithmetic mean of a population (μ) is one measure of central tendency and is given by:

$$\mu = \sum_{i=1}^{N} X_i / N \qquad (1.1)$$

where N is the number of observations, X_i is the set of measurements and $\sum$

is the symbol for summation. In other words, starting with the first value of X (i=1) add up all N values and divide the sum by N.

The mean of a sampled population (x) is given by:

$$\bar{x} = \sum X_i / N \qquad (1.1a)$$

where the sum is performed over the same range as in (1.1).

Exercise 1.1

(1) Compute the means of the 4 lithologies given in Table 1.1 - typically one is interested in summarizing the columns (variables) of the matrix.

(2) What meaning can you attach to the sums of the rows in Table 1.1? to the column sums? it would be easy to compute the means of the rows of the matrix. what would the mean of row one be measuring? if the units of the four variables were different would it make sense to compute the mean of a row of a matrix?

(3) If the row sums for a data matrix (like Table 1.1) are constant the data set is said to be **closed**. is Table 1.1 closed? if a data set is not closed it is **open**.

(4) if the mean of a column is 15' show that :

$$\sum_{i=1}^{5} \mu = 5*15 = 75 \quad \text{or, that} \quad \sum_{i=1}^{N} K = NK$$

where K is a constant

(5) if $X_i = 1,2,3,4$ and $Y_i = 6,7,8,9$ does

$$\sum_{i=1}^{4} (Y_i - X_i) = \sum_{i=1}^{4} Y_i - \sum_{i=1}^{4} X_i \quad ?$$

In addition to a measure of some sort of average, we often need to have a way to express the variability present in a set of observations. For example the range is one possible measure; the difference between the biggest and smallest values. This measure may not be useful for many studies as it is based solely on the extreme values.

Each value of a variable deviates from its mean (μ) by an amount e_i given by:

$$e_i = X_i - \mu \qquad (1.2)$$

The sum of these deviations ($\sum e_i$), however, is not a useful measure of variability because it is always zero:

$$\sum e_i = \sum (X_i - \mu)$$

$$= \sum X_i - \sum \mu$$

$$= N\mu - N\mu$$

$$= 0$$

(each sum is over the range i=1 to i=N) Follow through the steps and make sure that you understand the result.

The average of the sum of squares of the deviations is the preferred measure of sample variability. The sum of squares (SS) is given by:

$$SS = \sum_{i=1}^{N} (X_i - \mu)^2 \qquad 2.3$$

In this form the calculation is a two-step process. First, the mean is computed and then the sum of squares of the deviations can be computed. A more efficient formulation is given by:

$$SS = \sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2 / N \qquad 2.4$$

The variance of the target population (∂^2) is given by:

$$\partial^2 = SS/N \qquad 2.5$$

For the sampled population the variance (s^2) is given by:

$$s^2 = SS/(N-1) \qquad 2.6$$

Division by (N-1) instead of (N) is required in order to produce the best estimate of the target population variance from the sampled population variance and will be discussed in class.

Exercise 1.2

(1) Show that the two forms of the sums of squares of the deviations (2.3 and 2.4) are equivalent.

(2) Create the matrix of deviations from the mean for the data in Table 1.1; compute the variances of the four variables. compute the sums of the columns and rows of the deviations matrix. can you find any restriction on these sums; that is, is the sum always 0.0 or some other value.

(3) What are the units of variance of the quartz arenite thickness? The square root of the variance is the standard deviation (∂ or s) and is often used as a measure of variability. what are the units of the standard deviation?

(4) Compute the standard deviations for the four variables in Table 1.1

(5) The sum of the variances for a set of variables may be used as a measure of the information content of the data set. The contribution that each variable makes to the sum is given by the percentage of the total variance contributed by each variable. compute these percentages for each of the 4 variables. What must be true about the units of measurement in order for this to be a reasonable calculation?

(6) The coefficient of variation (C) is given as the ratio of the standard deviation to the mean. For the ith variable -

$$C_i = s_i / x_i \qquad (2.7)$$

Compute the coefficients of variation for the four variables. What are the units of the coefficients of variation. Sometimes it is helpful to think of the coefficient of variation as providing a measure of relative variability (relative to the mean) whereas the variance is a measure of absolute variability.

(7) It is difficult to compare the variances of variables that differ in units of measure. For example, if you have a collection of objects and measured the thickness in mm and the width in cm, how much larger would the following statistics of the thickness be as compared to the same statistics for the width?

a) the mean
b) the variance
c) the standard deviation
d) the coefficient of variation

(8) To rank order a list of values you place the largest first in the list, the second largest second in the list and so on. Rank order the sample variances and then rank order a the sample coefficients of variation; compare and contrast the two lists; what are these lists telling you?

GEOLOGICAL ANALYSIS LONG 2
MEASURES OF JOINT VARIATION

Statistics introduced in exercise 1 - mean, variance, standard deviation, etc. - serve to summarize the distribution of a given variable and are termed univariate statistics. Frequently, however, one is most concerned with the pairwise behavior of the variables being studied. For example, how does sandstone thickness vary with red mudstone thickness? As one increases, how does the other change (increase, decrease, stay the same, ...)?

The covariance (σ_{ij} or cov_{ij} - target population or sample respectively) is defined as a measure of the joint variation of the variables about their means:

$$cov_{AB} = SPROD_{AB}/(N-1) \qquad (2.1)$$

where SPROD (the sum of the product of the deviations) is given by the following expression.

$$SPROD_{AB} = \sum(A_i-A)(B_i-B) \qquad (2.2)$$

where the sum is from i=1 to N (number of pairs of values).
For computer aided computation, the following form of (2.2) is more efficient.

$$cov_{AB} = (\sum(A_i B_i)-((\sum A_i \sum B_i)/N))/(N-1) \qquad (2.2a)$$

If variable A equals variable B (that is, replace B by A in equation 2.2a) the covariance of A with itself (cov_{AA}) equals the variance of variable A. The most efficient way to display covariances is in matrix form. For example, if the data matrix consists of 4 variables (A, B, C, and D) the covariances are arranged as follows.

Variance/Covariance Matrix

	A	B	C	D
A	cov_{AA}	cov_{AB}	cov_{AC}	cov_{AD}
B	cov_{BA}	cov_{BB}	cov_{BC}	cov_{BD}
C	cov_{CA}	cov_{CB}	cov_{CC}	cov_{CD}
D	cov_{DA}	cov_{DB}	cov_{DC}	cov_{DD}

From equation 2.2a, it should be clear that cov_{AB} is equal to cov_{BA} as B times A is the same as A times B; that is, the order of multiplication does not matter. Therefore, the matrix given above is symmetric; element e_{ij} equals element e_{ji}. Also, this matrix is square; as many rows as columns. For square symmetric matrices the trace is the sum of the values lying on the main diagonal; the elements with the same row and column indices; e_{11}, e_{22}, etc. For the variance/covariance matrix the trace is equal to the sum of the variances of the variables. Covariances can be positive or negative. If the values of A and B increase together the covariance is positive (+). On the other hand, if B increases as A decreases the covariance is negative (-).

We will make use of the variance/covariance matrix as a measure of similarity in multivariate analyses later on in the course. One aspect of this matrix that should be noted at this time is the dimensions of the covariance; the product of the measures of the variables. Thus, the covariance between the weights of class members and their height in mm would have units of pounds·mm. Thus, like the variance, the covariance is weighted according to the magnitudes of the variabes.

Questions:

2.1 let X_i take on values of 5,5,5,5,5; what is the mean; the standard deviation.

2.2 by substitution, show that 2.2 is equivalent to 2.2a

2.3 using the deviation matrix prepared in the first long exercise, compute the matrix of covariances

2.4 compute the sums of the rows and columns of the covariance/variance matrix; are there any obvious restrictions on these sums; ie, do they sum to the same value or to zero, etc.

A measure of pairwise similarity that is dimensionless is the Pearson product moment correlation coefficient; r_{AB}.

$$r_{AB} = cov_{AB}/s_A s_B \qquad (2.3)$$

Note that the correlation of a variable with itself is +1.0. The correlation coefficient ranges from +1.0 to -1.0. The sign of r is the same as the sign of the covariance. The correlation provides a single number that measures the strength of the linear association that exists between a pair of variables. A coefficient of +1.0 indicates a perfect linear relationship with a + slope between the pair of variables whereas an r of 0.0 indicates a buckshot pattern. These measures of similarity are usually depicted as a square matrix.

Questions

2.1 - compute the matrix of correlations for the data set.

2.2 plot green mudstone versus quartz arenite; plot quartz arenite versus biomicrite and plot green mudstone versus red mudstone; compare the computed correlations with the correlations. does r convey the degree of linear association?

2.3 for M variables, the trace of the correlation coefficient is _____

Long Exercise 3
Percentage Formation

Form percentages for the lithology data set and compute the summary statistics; this time you can use FILE MASTER if you build the file. For the mean, variance and all the correlations prepare the following table in which you rank order the open statistics (largest listed first) and write in the values for the closed.

Means

OPEN	CLOSED
B - 50.2	Mean of %B
A - 45.3	Mean of %A

Compare the open and closed values (briefly)